Statistique et Aide à la Décision 2023

David Causeur

Analyse comparative des voeux confirmés sur Parcoursup en 2022 et 2023

Dans un contexte de forte évolution de la formation scientifique au lycée, On cherche à comparer les profils des élèves ayant confirmé leur voeu de candidater au cursus d'ingénieur en alimentation de l'Institut Agro Rennes Angers en 2022 et en 2023.

On dispose pour cela de données extraites de ParcourSup, donnant un grand nombre d'informations pour chacun.e de ces élèves candidat.e.s. On restreint la liste des élèves à celles et ceux en classe de terminale dans la filère générale :

```
postbac = read.table("https://dcauseur.netlify.app/teaching/data/postbac.txt",
                     header=TRUE,
                     stringsAsFactors = TRUE)
# Conversion de la variable Annee en variable catégorielle
postbac$Annee = factor(postbac$Annee)
# Classement des modalités de la variable Specialites (Autres en dernier)
specialites = levels(postbac$Specialites)
postbac$Specialites = ordered(postbac$Specialites,
                              levels=specialites[c(2,3,4,1)])
# Résumé des données
summary(postbac)
```

```
  Annee        Specialites       Genre
 2022:212   MATHS/PC :137   Féminin :223
 2023:175   MATHS/SVT:107   Masculin:164
            PC/SVT   :129
            Autres   : 14
```

Toutes les variables du tableau sont catégorielles. La variable `Spécialites` donne le choix de la doublette de spécialités choisie par chaque élève candidat.e.

A partir de ces données, on cherche en particulier à répondre à la question suivante : **la répartition des élèves candidat.e.s selon leurs choix de spécialités est-elle la même en 2022 et en 2023 ?**

Dans un premier temps, on calcule les proportions par année d'élèves ayant choisi chacune des doublettes de spécialités.

La table de contingence suivante contient toutes les informations permettant d'étudier le lien entre le choix des spécialités et l'année :

```
tab <- table(postbac$Annee,postbac$Specialites)
tab
```

```
      MATHS/PC MATHS/SVT PC/SVT Autres
  2022       66        54     82     10
  2023       71        53     47      4
```

La fonction `rowPercents` du package `RcmdrMisc` permet de transformer cette table de contingence de sorte que chaque ligne contienne les proportions des élèves ayant choisi chacune des doublettes de spécialités par année :

```
prop <- RcmdrMisc::rowPercents(tab)
prop
```

```
      MATHS/PC MATHS/SVT PC/SVT Autres Total Count
  2022     31.1      25.5   38.7    4.7 100.0   212
  2023     40.6      30.3   26.9    2.3 100.1   175
```

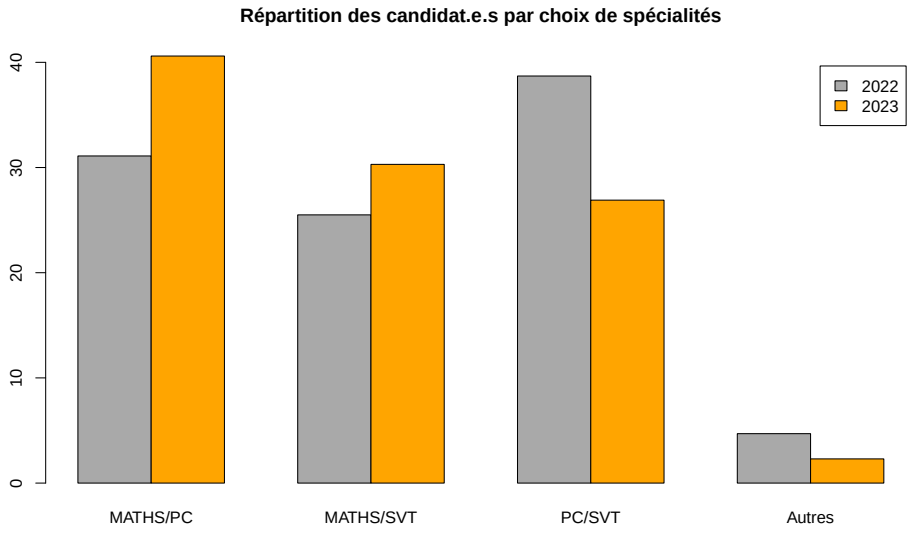
De manière équivalente, on peut utiliser le package `tidyverse` pour extraire les mêmes informations :

```
prop_tidy <- postbac %>% select(Annee,Specialites) %>%
  group_by(Annee) %>%
  count(Specialites) %>%
  mutate(Pourcentages=100*n/sum(n))
prop_tidy
```

```
# A tibble: 8 x 4
# Groups:   Annee [2]
  Annee Specialites     n Pourcentages
  <fct> <ord>       <int>        <dbl>
1 2022  MATHS/PC       66        31.1
2 2022  MATHS/SVT      54        25.5
3 2022  PC/SVT         82        38.7
4 2022  Autres         10         4.72
5 2023  MATHS/PC       71        40.6
6 2023  MATHS/SVT      53        30.3
7 2023  PC/SVT         47        26.9
8 2023  Autres          4         2.29
```

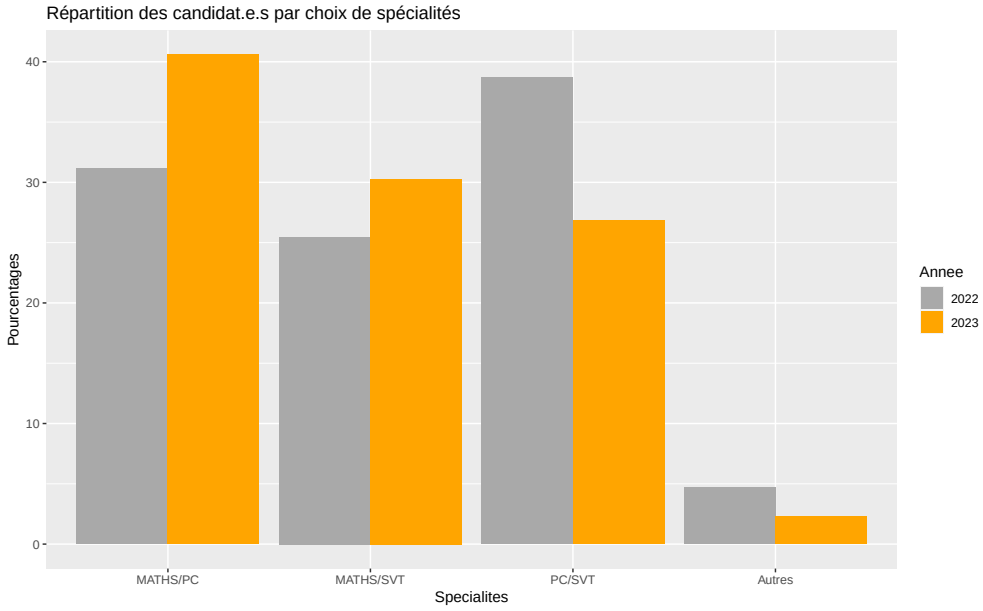
On peut apprécier visuellement les différences de répartition des choix de spécialités entre année à partir du graphe suivant :

```
barplot(prop[,1:4],beside=TRUE,
        col=c("darkgray","orange"),
        legend.text=c("2022","2023"),
        main="Répartition des candidat.e.s par choix de spécialités")
```



On peut générer le même type de graphique en utilisant le package `tidyverse` :

```
prop_tidy %>% ggplot() +
  aes(fill=Annee,y=Pourcentages,x=Specialites) +
  geom_bar(position="dodge",stat="identity") +
  scale_fill_manual(values=c("darkgray", "orange")) +
  ggtitle("Répartition des candidat.e.s par choix de spécialités")
```



Pour tester si la différence de répartition des élèves par choix de spécialités entre 2022 et 2023 est significative, on utilise un test du χ^2 de Pearson :

```
chisq.test(tab)
```

```
    Pearson's Chi-squared test

data:  tab
X-squared = 8.802, df = 3, p-value = 0.032
```

Au seuil de 5%, on décide donc que les répartitions des élèves par choix de spécialités en 2022 et 2023 sont significativement différentes. Afin d'aller plus loin dans l'interprétation de ce résultat, on analyse les contributions de chaque cellule du tableau de contingence à la statistique du test :

```
chisq.test(tab)$residuals
```

```
       MATHS/PC MATHS/SVT    PC/SVT    Autres
2022 -1.044558 -0.602790  1.348188  0.841626
2023  1.149692  0.663460 -1.483882 -0.926335
```

On observe que la principale différence entre 2022 et 2023 est la diminution de la proportion d'élèves ayant choisi les spécialités Physique-Chimie et Sciences de la vie et de la terre.

On peut exprimer la contribution de cette différence à la statistique de test par la transformation en pourcentages suivante du tableau des carrés des résidus :

```
100*chisq.test(tab)$residuals^2/chisq.test(tab)$statistic
```

```
      MATHS/PC MATHS/SVT   PC/SVT   Autres
2022 12.39553   4.12793 20.64909  8.04708
2023 15.01630   5.00069 25.01490  9.74847
```

La différence observée ci-dessus et concernant le choix des spécialités Physique-Chimie et Sciences de la vie et de la terre contribue à 45.65% à la statistique de test.